ALIGNED GEOMETRY LATENT SPACE FOR IMPROVED WORLD CONSISTENCY IN JOINT VIDEO-GEOMETRY MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

While jointly modeling video with compressed geometry latents has been shown to effectively improve geometric consistency in video generation, such methods typically suffer from slow convergence or training instability. Our analysis suggests that this issue is closely related to the misalignment between compressed geometry latents and pretrained video models: *higher reconstruction fidelity does not necessarily make the latents more effective for video models to understand and utilize.* Motivated by this, we propose geometry latent alignment (GeoLA), a post-training framework that directly leverages the video generator to supervise geometry compression and produce a well-aligned geometry latent. Specifically, we use geometry latents compressed by a geometry adapter as conditions for video denoising, and optimize the adapter with the diffusion loss. This allows the video model to implicitly guide the adapter to retain the geometric information most useful for denoising. We then freeze the aligned adapter for joint video-geometry modeling, followed by dual-stream distribution matching distillation for low-latency joint generation. To support training and evaluation, we further introduce GeoVid-80K, a dataset containing over 80,000 video clips with rich geometric variations. Experiments show that GeoLA converges faster during joint training and generates videos with improved geometric consistency.

1 INTRODUCTION

Video generation models are becoming an important paradigm for modeling dynamic visual environments (Sun et al., 2025; Huang et al., 2025a; Chen et al., 2025; HY-World et al., 2026; Dai et al., 2025; Wei et al., 2026). As they are increasingly used for applications such as autonomous driving (Hu et al., 2023; Wang et al., 2024), embodied intelligence, and interactive world simulation, these models must provide not only visual realism but also geometric consistency across different viewpoints. Although recent large-scale video diffusion models (Wan et al., 2025; Kong et al., 2024; Hong et al., 2023; Xiang et al., 2025) can generate highly realistic videos, they still struggle to maintain consistent geometry under large viewpoint changes or over long generation horizons (Bhowmik et al., 2025; Xue et al., 2025; Park et al., 2025).

Joint geometry–video modeling provides a promising post-training approach for improving geometric consistency (Byung-Ki et al., 2025; Lin et al., 2026; Chefer et al., 2025; Xi et al., 2026; Bai et al., 2026). By modeling geometry and appearance within a shared denoising process, the model can learn the geometric structure underlying visual content and better preserve it during generation.

Motivated by the strong geometric perception capabilities of large reconstruction models (LRMs) (Wang et al., 2025c;b; Lin et al., 2025), recent works leverage intermediate features from LRMs for joint video–geometry modeling and compress these features to alleviate the difficulty of diffusion models in learning high-dimensional feature spaces (Li & He, 2026). As shown in Figure 1, DreamWorld (Tan et al., 2026) uses principal component analysis (PCA) for dimensionality reduction, while Gen3R (Huang et al., 2026a) trains an additional geometry VAE with a reconstruction objective and KL loss for compression. While effective, they often suffer from slow convergence during joint training and require carefully designed training strategies (Tan et al., 2026) to mitigate optimization instability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

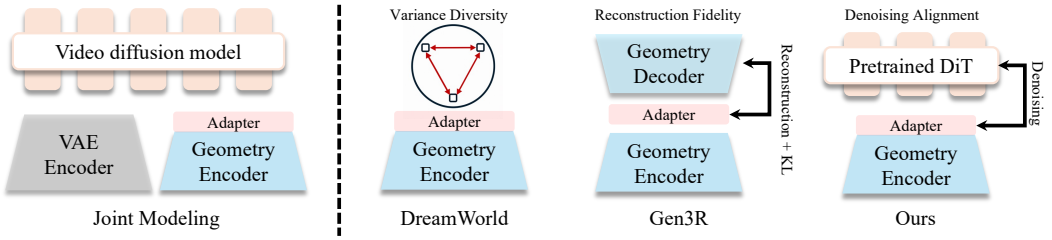


Figure 1: **Comparison of geometry compression for joint video–geometry modeling.** Existing methods use PCA-based projection, or reconstruction-based compression, whereas our method learns the geometry latent under the supervision of a pretrained video denoising model.

To systematically investigate the source of these training difficulties, we conduct analysis experiments on different geometry compression strategies (Tan et al., 2026; Huang et al., 2026a) in both controllable video generation and joint video–geometry modeling. Existing compression methods are typically optimized to reconstruct the original geometry features, raising the question of *whether reconstruction fidelity reflects the suitability of the resulting latents for video diffusion.* Our experimental results show that latents with higher reconstruction fidelity yield weaker geometric control in controllable generation and slower convergence in joint modeling. These findings suggest that reconstruction fidelity may not be an appropriate optimization objective for geometry compression.

Motivated by this, we propose GeoLA, a novel framework for joint geometry-video modeling. The core idea is to directly use the diffusion objective of the pretrained video model to compress and align the geometry latents. Specially, in the first stage, a lightweight adapter compresses LRM features into a low-dimensional latent, which are then used to condition the denoising process of the pretrained video model. We then directly optimize the geometry adapter with the video model's flow-matching loss, encouraging it to preserve the geometric features most useful for video denoising. In the second stage, we freeze the learned geometry space and jointly model the geometry and video latents with flow matching. In the third stage, we introduce a dual-stream distribution-matching distillation scheme, distilling the joint generative model into an efficient few-step generator that enables low-latency video generation while preserving strong geometric consistency.

To support training and evaluation, we further construct *GeoVid-80K*, a video dataset with paired appearance and geometry latents. GeoVid-80K contains over 80,000 video clips spanning diverse indoor and outdoor environments, with high-quality camera and subject motion.

2 Problem Analysis

To investigate the source of the optimization difficulty in joint geometry-video modeling, we conduct controlled experiments on representative geometry compression strategies, including PCA-based compression (Tan et al., 2026) and its alignment variant, as well as reconstruction-based compression (Kingma & Welling, 2013) and its alignment variant (Huang et al., 2026a).

Specifically, we conduct two complementary experiments to evaluate the compressed geometry latents in downstream video generation. First, we use the geometry latent as a condition for geometry-controlled video generation, where control accuracy measures whether the latents can be effectively understood and utilized by the video model. Second, we use the geometry latent as a generation target for joint geometry-video modeling, where training convergence reflects whether the latents can be effectively learned and modeled. The experiments lead to the following observations.

Reconstruction fidelity does not guarantee effective geometric constraints on video generation. As shown in Fig. 2(a), **Rec.** achieves lower reconstruction error than **Rec.+KL**. However, when used to condition the pretrained video generator, **Rec.+KL** consistently yields lower conditional video loss (Fig. 2(b)), suggesting that the generator can utilize its geometry latents more effectively. This advantage also translates into better geometric control: at later checkpoints, **Rec.+KL** achieves lower depth AbsRel and higher δ_1 (Fig. 2(d,e)), indicating closer agreement between generated and reference videos.

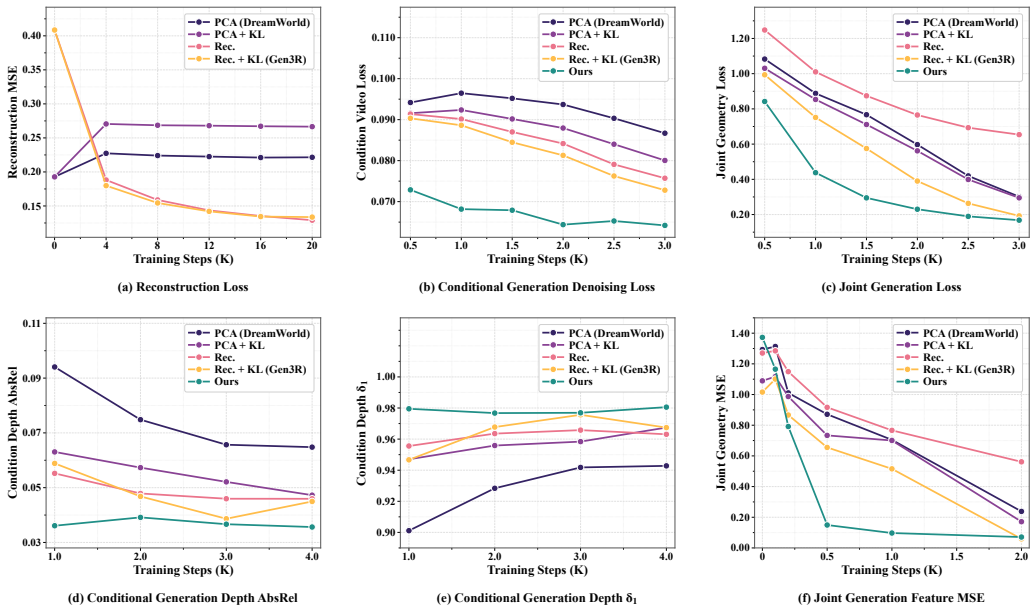


Figure 2: **Analysis of geometry latent compression.** Higher reconstruction fidelity does not necessarily lead to stronger geometric control or faster convergence in joint video–geometry modeling.

Reconstruction fidelity does not guarantee efficient learnability by video models. For all latent-space in Fig. 2(c,f), we normalize each geometry latent using per-channel mean and standard deviation estimated over the training set, following the standard latent normalization used in VAE-based generative models. As shown in Fig. 2(c), **PCA+KL** and **Rec.+KL** converge substantially faster than **PCA** and **Rec.**, respectively, during joint geometry–video modeling, despite not necessarily achieving better geometry reconstruction. Beyond the training loss, both aligned variants also exhibit lower MSE between predicted geometry latents and those re-encoded from the generated videos (Fig. 2(f)), indicating that the model learns the geometry–video joint distribution more efficiently.

Motivation. These observations suggest that higher reconstruction fidelity does not necessarily translate into better downstream video generation. This naturally motivates a different compression objective: rather than preserving geometry solely for accurate reconstruction, can we let the diffusion model itself determine which geometric components are most useful for downstream denoising ?

3 Method

3.1 Preliminary

Distribution matching distillation. Distribution matching distillation (DMD) (Yin et al., 2024; Huang et al., 2025b) trains a few-step generator G_θ by matching its generated distribution to the real data distribution through score differences. Given Gaussian input noise $\boldsymbol{\eta}$, the generator produces $\hat{\mathbf{x}}_0 = G_\theta(\boldsymbol{\eta})$, which is perturbed as $\hat{\mathbf{x}}_t = F(\hat{\mathbf{x}}_0, \boldsymbol{\epsilon}, t)$. The generator is updated by

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{t,\boldsymbol{\eta},\boldsymbol{\epsilon}} \left[\Delta s(\hat{\mathbf{x}}_t, t) \frac{\partial G_\theta(\boldsymbol{\eta})}{\partial \theta} \right], \tag{1}$$

where $\Delta s(\hat{\mathbf{x}}_t, t) = s_{\text{real}}(\hat{\mathbf{x}}_t, t) - s_{\text{fake}}(\hat{\mathbf{x}}_t, t)$. Here s_{real} is provided by a large-scale frozen teacher model, while s_{fake} is estimated by a small-scale diffusion model trained on samples from G_θ.

3.2 Stage 1: Compress and Align Geometry Latent to Pretrained Video model

Raw features $\mathbf{G}$ from geometry foundation models differ substantially from video latents in both spatial resolution and channel dimensionality. Direct joint diffusion in this high-dimensional space

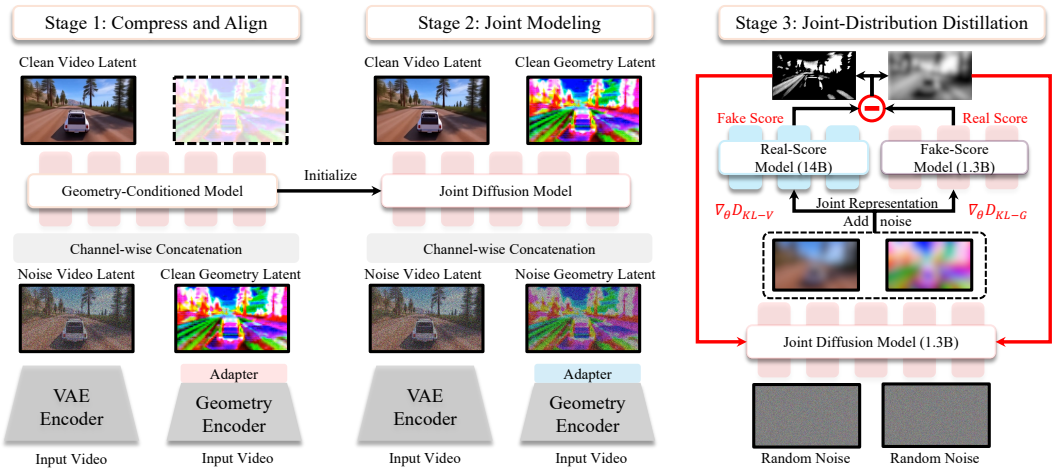


Figure 3: **Overall pipeline of GeoLA.** The framework consists of three stages: (1) **Compress and Alignment**: Align geometry latents with the pretrained video diffusion; (2) **Joint Modeling**: learning a unified video–geometry diffusion space; and (3) **Distillation**: transferring the joint distribution to a compact generator for efficient and 3D-consistent video generation.

is not only computationally expensive but also difficult to optimize effectively. We therefore introduce a lightweight geometry adapter $\mathcal{A}_\psi$ that compresses the geometry features while aligning the resulting latent space with the pretrained video model through the video flow-matching objective.

Adapter Architecture. The adapter $\mathcal{A}_\psi$ resizes and projects raw geometry features into a low-dimensional space aligned with the video latent grid:

$$\mathbf{g}_\psi = \mathcal{A}_\psi(\mathbf{G}) = \mathcal{N}(\mathcal{P}_\psi(\mathcal{R}(\mathbf{G}))) \in \mathbb{R}^{B\times T\times C_g\times H\times W}, \tag{2}$$

where $\mathcal{R}$ resizes features to target resolution (H, W), $\mathcal{P}_\psi$ is a token-wise MLP projecting to a tight bottleneck dimension C_g, and $\mathcal{N}$ denotes non-affine batch normalization. The tight channel capacity forces $\mathcal{A}_\psi$ to discard irrelevancies and compress structural essentials.

Compression with Flow-Matching Loss. To align geometry compression directly with the pretrained video model, we optimize the geometry adapter $\mathcal{A}\psi$ through the video flow-matching objective. Specifically, the compressed geometry latent $\mathbf{g}\psi$ is incorporated into the pretrained video Diffusion Transformer (DiT) as an additional conditional input, allowing gradients from the video denoising objective to directly supervise the compression process end-to-end. To preserve the pretrained video prior at initialization, we expand the input projection with zero-initialized channels, i.e., $\widetilde{\mathbf{W}}\mathrm{in} = [\mathbf{W}\mathrm{in}, \mathbf{0}g]$. Then, the adapter $\mathcal{A}_\psi$ is updated end-to-end via the conditional video flow-matching loss:

$$\mathcal{L}_{\text{repr}} = \mathbb{E}_{k,\mathbf{z}_0,\mathbf{G},\boldsymbol{\epsilon}_z,\rho,\boldsymbol{\epsilon}_g}\left[w(\sigma_k)\left\|v_\theta^{\text{cond}}(\mathbf{z}^+_{\sigma_k}, \mathbf{g}\psi, \mathbf{z}^c_0, \sigma_k) - \mathbf{u}^+_z\right\|_2^2\right], \tag{3}$$

where $\mathbf{z}^+_{\sigma_k}$ is the noised target video latent, $\mathbf{z}^c_0$ is the clean first-frame context, and $\mathbf{u}^+_z = \boldsymbol{\epsilon}_z - \mathbf{z}^+_0$ is the ground-truth velocity. Gradient feedback from $\mathcal{L}_{\text{repr}}$ enables the adapter to adaptively select and compress geometry information most useful for video denoising, while shaping the resulting latent space to better align with the pretrained video model.

3.3 Stage 2: Joint Flow Matching in a Fixed Geometry Latent Space

With the geometry latent space aligned to the pretrained video model, we freeze the adapter parameters $\psi^\star$ to anchor the geometry latent target: $\mathbf{g}_0 = \mathrm{sg}[\mathcal{A}_{\psi^\star}(\mathbf{G})]$. Fixing this target prevents the geometry representation from drifting during generative training, thereby providing a stable latent space for learning the coupled video–geometry distribution.

Table 1: **Overview of the composition and distribution of GeoVid-80K.**

Source	Scene regime	Data type	Primary coverage	Clips	Ratio
RealEstate10K	Static	Real	Indoor and residential scenes	33,348	39.29%
MiraData	Dynamic	Real + Synthetic	Open-domain object and human motion	17,679	20.83%
SpatialVID	Dynamic	Real	Diverse scenes and camera trajectories	15,006	17.68%
Private collection	Dynamic	Real	Tracking, interaction, and aerial footage	18,844	22.20%
GeoVid-80K	**Mixed**	**Mixed**	**Open-domain world videos**	**84,877**	**100.00%**

Conditioning and Noise Injection. We use task-specific conditioning schemes for text-to-video and image-to-video generation. For text-to-video generation, all T video and geometry frames are jointly noised using the same timestep σ_k. For image-to-video generation, the first video–geometry frame pair $(\mathbf{z}_0^c, \mathbf{g}_0^c)$ is kept clean as conditioning context, while the remaining $T-1$ frames are noised as generation targets. In both settings, the noised target latents are constructed as

$$\mathbf{z}_{\sigma_k}^+ = (1-\sigma_k)\mathbf{z}_0^+ + \sigma_k \boldsymbol{\epsilon} * z, \qquad \mathbf{g}_{\sigma_k}^+ = (1-\sigma_k)\mathbf{g}_0^+ + \sigma_k \boldsymbol{\epsilon}_g, \tag{4}$$

where $\boldsymbol{\epsilon}_z$ and $\boldsymbol{\epsilon}_g$ are independently sampled Gaussian noise. Here, $\mathbf{z}_0^+$ and $\mathbf{g}_0^+$ denote all T frames for text-to-video generation and the $T-1$ target frames excluding the clean context frame for image-to-video generation.

Dual-Head Architecture and Asymmetric Gradient Flow. The joint network takes concatenated target latents $\mathbf{y}_{\sigma_k}^+ = [\mathbf{z}_{\sigma_k}^+, \mathbf{g}_{\sigma_k}^+]$ and clean context $\mathbf{y}_0^c = [\mathbf{z}_0^c, \mathbf{g}_0^c]$. To encourage shared 3D representation learning without compromising video quality, we attach a newly initialized geometry head $\mathcal{O}_g$ to an intermediate layer $\ell < L$, while retaining the video head $\mathcal{O}_z$ at the final layer L:

$$\widehat{\mathbf{u}}_{z,\phi} = \mathcal{O}_z\big(\mathbf{h}_\phi^L(\mathbf{y}_{\sigma_k}^+, \mathbf{y}_0^c, \sigma_k)\big), \qquad \widehat{\mathbf{u}}_{g,\phi}^\ell = \mathcal{O}_g\big(\mathbf{h}_\phi^\ell(\mathbf{y}_{\sigma_k}^+, \mathbf{y}_0^c, \sigma_k)\big). \tag{5}$$

The joint model is optimized using a combined velocity prediction loss:

$$\mathcal{L}_{\text{joint}} = \mathbb{E}\big[w(\sigma_k)\,\big(\text{MSE}(\widehat{\mathbf{u}}_{z,\phi}, \mathbf{u}_z^+) + \text{MSE}(\widehat{\mathbf{u}}_{g,\phi}^\ell, \mathbf{u}_g^+)\big)\big]. \tag{6}$$

This design allows lower-to-mid transformer layers to master joint spatio-temporal structure, leaving deeper layers to concentrate on high-fidelity video texture synthesis.

3.4 Stage 3: Joint-Distribution Distillation

The joint diffusion model improves geometry-video consistency but remains costly due to iterative denoising. We therefore distill its joint distribution into a few-step generator G_ω. Given Gaussian noise $\boldsymbol{\eta}$ and condition $\mathbf{c}$, the student directly predicts $\hat{\mathbf{y}}_0 = G_\omega(\boldsymbol{\eta}, \mathbf{c}) = [\hat{\mathbf{z}}_0, \hat{\mathbf{g}}_0]$, where $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{g}}_0$ are the generated video and geometry latents. The generated joint latent is perturbed as $\hat{\mathbf{y}}_t = F(\hat{\mathbf{y}}_0, \boldsymbol{\epsilon}_y, t)$. Following DMD, the student is updated by

$$\nabla_\omega \mathcal{L}_{\text{JDD}} = -\mathbb{E}_{t,\boldsymbol{\eta},\boldsymbol{\epsilon}_y}\left[w(t)\Delta s^j(\hat{\mathbf{y}}_t, t, \mathbf{c})\frac{\partial G_\omega(\boldsymbol{\eta}, \mathbf{c})}{\partial \omega}\right], \tag{7}$$

where $\Delta s^j(\hat{\mathbf{y}}_t, t, \mathbf{c}) = s_{\text{T}}^j(\hat{\mathbf{y}}_t, t, \mathbf{c}) - s_{\text{G}}^j(\hat{\mathbf{y}}_t, t, \mathbf{c})$. Here s_{T}^j is estimated by the frozen geometry-video joint teacher, while s_{G}^j is estimated by a fake-score model trained on samples from G_ω. This objective transfers the teacher's joint score field to the student, allowing the compact generator to preserve both visual fidelity and geometry-video consistency.

During inference, G_ω outputs $\hat{y}_0 = [\hat{z}_0, \hat{g}_0]$, but only the video latent $\hat{z}_0$ is decoded by the video VAE. After distillation, the generated video latent itself carries the distilled geometry-aware prior.

4 GeoVid-80K Dataset

To generalize geometry-aware video generation across diverse open-world scenes, we construct GeoVid-80K, a curated dataset comprising 84,877 video clips with aligned geometry supervision.

Data sources. GeoVid-80K combines four data sources, as summarized in Table 1. RealEstate10K (Zhou et al., 2018) provides static indoor scenes with substantial camera motion.

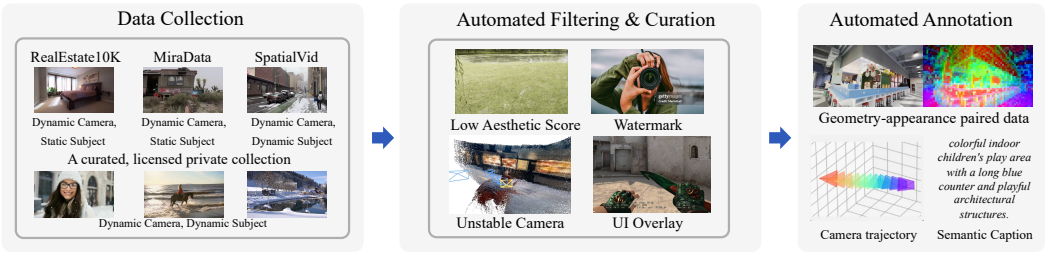


Figure 4: **Overview of the GeoVid-80K data collection, curation and annotation pipeline.**

Table 2: **Quantitative Comparison on Text-to-Video Quality, 3D Geometry, and Consistency.** indicates the best performance, the second best, and the third best.

Method	NFE ↓	VBench Metrics					3D Reconstruction Error			3D Consistency Metrics	
		Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑	Aesthetic Quality ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Epipolar Error ↓	Inlier Rate ↑
OmniVDiff	50 × 2	0.9309	0.9230	0.9875	0.6135	0.5330	19.5630	0.6157	0.4169	9.2397	0.6133
GeoVideo	50 × 2	0.9182	0.8839	0.9646	0.6334	0.5340	18.9986	0.6349	0.3975	33.7536	0.3033
VideoGPA	50 × 2	0.8972	0.8155	0.9647	0.6314	0.5740	17.0392	0.5353	0.4061	11.1430	0.5866
GeoLA-T2V	5	0.9305	0.9530	0.9910	0.7383	0.5988	21.4994	0.7082	0.2987	7.2780	0.7068

High-resolution subsets of SpatialVID (Wang et al., 2025a) and MiraData (Ju et al., 2024) contribute dynamic open-domain videos. A curated, licensed private collection further adds complex camera and object motions, including tracking shots and object interactions.

Visual and technical filtering. To ensure basic visual and technical quality, we use the LAION aesthetic predictor (Schuhmann et al., 2022) to filter low-quality content and PaddleOCR to remove clips dominated by subtitles, logos, or watermarks.

Motion and geometry filtering. To ensure that the datasets provide clear and reliable geometric supervision, we further filter the clips based on their estimated depth and camera motion using DepthAnything3 (Lin et al., 2025). We remove videos with little viewpoint or subject motion, unstable camera movement, and temporally inconsistent depth. For dynamic scenes, we additionally require the subject motion to be sufficiently clear while the background remains relatively stable.

Manual Curation. To further improve the quality of the dataset, we additionally conduct 120 person-hours of manual inspection to identify and remove residual low-quality samples. The inspection primarily focuses on visual quality, temporal coherence, camera-motion stability, and whether subject motion is plausible and clearly observable. The final dataset contains 84,877 clips and approximately 6.88 million RGB frames with synchronized geometry annotations.

Camera Trajectory Annotation. We use DepthAnything3 (Lin et al., 2025) to estimate the camera trajectory of each video clip. The estimated trajectory records the frame-wise camera pose and provides explicit camera-motion annotations for subsequent training and evaluation.

Semantic Description Caption. To provide rich semantic annotations, we employ Qwen-2.5-VL-72B (Bai et al., 2025b) to generate detailed textual descriptions for each video clip. The generated captions summarize the scene content, major subjects, their appearance and motion, as well as relevant camera motion and interactions. These descriptions complement the geometric annotations with high-level semantic information.

5 Experiments

Baselines. For T2V, we compare GeoLA with OmniVDiff (Xi et al., 2026), GeoVideo (Bai et al., 2026), and VideoGPA (Du et al., 2026). For I2V, we compare with GeometryForcing (Wu et al., 2025), ViewCrafter (Yu et al., 2024b), Gen3C (Ren et al., 2025), Gen3R (Huang et al., 2026a). For all baselines, we use the official open-source implementations and released model weights.

Implementation Details. We use Wan-1.3B (Wan et al., 2025) as the video diffusion backbone and DepthAnything3 (Lin et al., 2025) to extract geometry features. For distillation, We use Wan-14B as

Table 3: **Quantitative Comparison on Image-to-Video Quality, 3D Geometry, and Consistency.** indicates the best performance, the second best, and the third best.

Method	NFE ↓	VBench Metrics					3D Reconstruction Error			3D Consistency Metrics	
		Background Consistency ↑	Subject Consistency ↑	Motion Smoothness ↑	Imaging Quality ↑	Aesthetic Quality ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Epipolar Error ↓	Inlier Rate ↑
GeometryForcing	50 × 2	0.9747	0.9653	0.9870	0.6918	0.4875	17.9446	0.6417	0.4403	-	-
ViewCrafter	50 × 2	0.9717	0.9621	0.9879	0.6898	0.5672	18.7581	0.6651	0.3896	21.9872	0.3301
Gen3C	35 × 2	0.9684	0.9695	0.9955	0.7250	0.5673	19.9375	0.6755	0.3630	5.7102	0.7578
Gen3R	50 × 2	0.9821	0.9675	0.9935	0.6918	0.5832	19.3977	0.6864	0.4104	5.5425	0.7708
GeoLA-I2V	5	0.9573	0.9684	0.9936	0.7143	0.5483	21.1826	0.7275	0.3109	5.3582	0.7788



Figure 5: **Qualitative Comparison of T2V Generation.** GeoLA produces more stable layouts and fewer geometric distortions on diverse indoor and outdoor scenes.

the teacher model. By default, we select features from LRM encoder layer 20 before geometry-latent adaptation. All training stages use AdamW with a global batch size of 48. More implementation details are provided in the supplementary material.

Evaluation Protocol. For T2V, we evaluate on 100 prompts covering diverse scenes, motions, and camera movements; for I2V, we use 100 RealEstate10K videos excluded from training. We assess generated videos from three aspects: video quality using VBench (Huang et al., 2024); 3D reconstruction consistency by reconstructing each video with 3DGS and comparing re-rendered and generated frames using PSNR (Horé & Ziou, 2010), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018); and epipolar consistency via SIFT-based cross-frame matching (Kupyn et al., 2025), measured by Epipolar Error and Inlier Rate. We also report the number of function evaluations (NFE) as a proxy for inference speed.

5.1 Main Results

Text-to-Video Quantitative Results. As shown in Table 2, GeoLA achieves a strong balance between visual quality and geometric consistency. On VBench, it attains the best subject consistency, motion smoothness, and imaging quality. More importantly, GeoLA consistently outperforms all baselines on the 3D consistency metrics, including PSNR, SSIM, LPIPS, and epipolar error. The improvement over both conventional 2D video generators and methods with explicit geometry supervision indicates that jointly learning video appearance and geometry is more effective than imposing geometry as an external constraint.

Text-to-Video Qualitative Results. Figure 5 further demonstrates the advantage of GeoLA in preserving 3D structure over time. Across diverse scenes and camera motions, GeoLA maintains more consistent object shape, scene layout, and viewpoint-dependent appearance. In contrast, competing methods are more prone to structural deformation, identity drift, and spatial inconsistency across frames. These results show that GeoLA produces videos with more stable scene geometry under dynamic viewpoint changes.

Image-to-Video Quantitative Results. As shown in Table 3, GeoLA-I2V achieves the best performance across all 3D reconstruction and epipolar consistency metrics. Compared with the larger-scale Gen3C, it maintains comparable VBench performance while consistently improving PSNR, SSIM, LPIPS, epipolar error, and inlier rate. This demonstrates that GeoLA-I2V improves geomet-

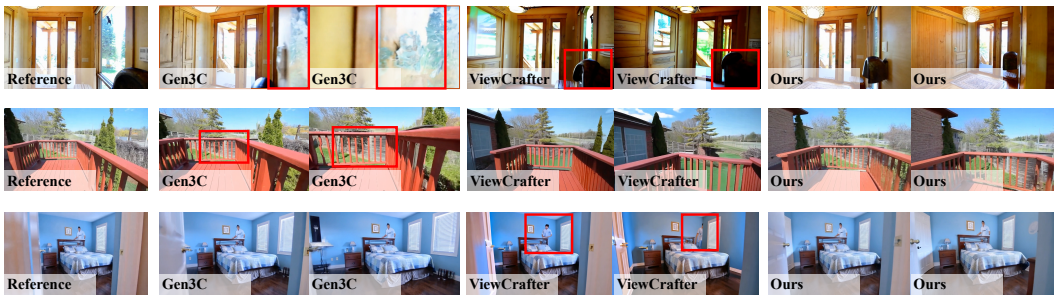


Figure 6: **Qualitative Comparison of I2V Generation.** Methods that rely on explicit reconstruction results as generation guidance tend to propagate depth estimation errors into the video generator.

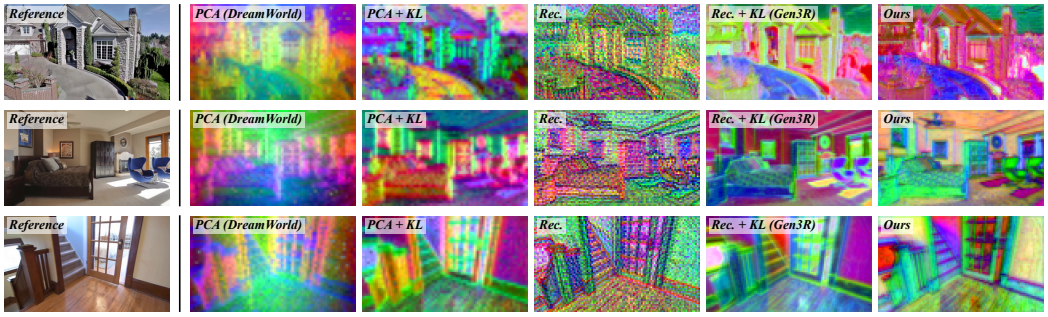


Figure 7: **Visualization of geometry latents using different compression strategies.**

ric consistency while maintaining competitive visual quality. GeometryForcing is excluded from the epipolar evaluation because its native output resolution is insufficient for reliable SIFT matching.

Image-to-Video Qualitative Results. Figure 6 shows that GeoLA-I2V maintains more coherent scene structure under large camera motion. Explicit geometry-guided methods are sensitive to errors in reconstructed geometry, which become more pronounced as the target viewpoint moves farther from the reference image and can lead to structural distortion, missing content, and cross-view inconsistency. In contrast, GeoLA-I2V avoids relying on explicit reconstructed intermediates and produces more stable novel views across challenging camera trajectories.

5.2 Ablations and Analysis

Unless otherwise specified, we conduct all ablation studies under the text-to-video setting to ensure consistent and controlled comparisons across different design choices. All variants use the same training data, joint modeling architecture, and evaluation protocol.

Geometry Compression and Alignment. To evaluate different geometry compression strategies, we compare four variants: PCA projection from DreamWorld (Tan et al., 2026), reconstruction and reconstruction+KL objectives from Gen3R (Huang et al., 2026a), and our learnable adapter. For a controlled comparison, all variants use features from the same geometry reconstruction model, while keeping the training set, model architecture, and feature fusion scheme identical, with geometry latents concatenated along the channel dimension. Notably, the other compression strategies do not update the DiT during the compression stage. To isolate the benefits of end-to-end adapter optimization with the diffusion loss from those of additional DiT training, we also perform DiT warm-up fine-tuning for each baseline before the joint modeling experiments. During this warm-up, the DiT is trained for conditional video generation using the compressed geometry latents as conditioning, and the resulting weights are then used to initialize the joint model. As shown in Table 4, our method achieves the strongest overall performance, with competitive video generation quality and improved

Table 4: **Ablation on alignment strategy.**

Method	VBench Metrics					3D Consistency	
	BC ↑	SC ↑	MS ↑	IQ ↑	AQ ↑	EE ↓	IR ↑
DreanWorld	0.9086	0.8240	0.9652	0.4305	0.4095	29.9325	0.2216
GeometryVAE	0.9093	0.9272	0.9859	0.7029	0.5349	12.6948	0.4974
Gen3R	0.9255	0.9517	0.9903	0.7001	0.5362	10.6972	0.6732
Ours	0.9317	0.9488	0.9891	0.7234	0.6189	9.4925	0.6737

geometric consistency. Figure 7 further shows that PCA produces less distinct structural boundaries, while reconstruction-based methods introduce substantial high-frequency feature noise. In contrast, our adapter better preserves fine-grained geometric structures, indicating that it learns a generation-compatible geometry representation rather than merely compressing the features.

Joint Geometry-Video Modeling. We compare standard video-only post-training (SFT) with our geometry-video joint post-training using the same training data without any distillation for a fair comparison. As shown in Table 5, joint geometry-video modeling consistently improves geometric consistency while maintaining strong video generation quality, demonstrating the benefit of jointly learning visual dynamics and 3D structure.

Table 5: **Ablation on joint modeling strategy.**

Method	VBench Metrics					3D Consistency	
	BC ↑	SC ↑	MS ↑	IQ ↑	AQ ↑	EE ↓	IR ↑
Wan-SFT	0.9086	0.9000	0.9712	0.6659	0.5633	19.9983	0.3935
Joint	0.9317	0.9488	0.9891	0.7234	0.6189	9.4925	0.6737

Table 6: **Ablation on distillation strategy.**

Method	VBench Metrics					3D Consistency	
	BC ↑	SC ↑	MS ↑	IQ ↑	AQ ↑	EE ↓	IR ↑
DMD	0.9209	0.8931	0.9718	0.7160	0.6103	18.2811	0.4110
Ours	0.9305	0.9530	0.9910	0.7383	0.5988	7.2780	0.7068

Distillation Strategy. We compare standard Video-DMD distillation with our joint video-geometry distillation strategy. As shown in Table 6, our method improves overall video quality and geometric consistency, with particularly notable gains under dynamic camera motion. While Video DMD effectively preserves the video generation capability of the teacher, joint score distillation further transfers the geometry-aware joint prior to the student, leading to more stable 3D structure and improved epipolar consistency while maintaining strong visual quality.

6 Related Work

Alignment with Vision Foundation Model (VFM). Pretrained vision models (Oquab et al., 2024) are widely used to supervise the internal representations of generative models. REPA (Yu et al., 2024a) aligns intermediate diffusion features from noisy inputs with clean-image features from a frozen visual encoder, improving training efficiency and generation quality. Related video methods similarly align denoiser features with representations from pretrained geometric foundation models to inject spatial structure (Wu et al., 2025). In these approaches, foundation-model features are fixed supervision targets. GeoLA instead learns the feature space itself: it constructs a compact geometry representation whose bottleneck is optimized for predictive utility in video denoising.

Joint Modeling with VFM Representations. Other methods model foundation-model features jointly with appearance as auxiliary generative variables (Huang et al., 2026b; Xie et al., 2026). DreamWorld jointly predicts video latents and compressed features from multiple foundation models to capture semantic, geometric, and temporal information (Tan et al., 2026), while Gen3R compresses features from a pretrained 3D reconstruction model into geometry latents and jointly generates them with appearance latents (Huang et al., 2026a). However, these compact feature spaces are defined by fixed statistical projections or reconstruction-oriented objectives. GeoLA instead first learns which geometric information should be preserved using video denoising supervision, then freezes this geometry space and models its joint distribution with video through diffusion.

7 Conclusion

We presented **GeoLA**, a post-training framework for improving geometric consistency in video generation through geometry latent alignment. Our analysis shows that higher reconstruction fidelity of compressed geometry features does not necessarily lead to better geometric control or faster convergence in joint video–geometry modeling. Based on this observation, GeoLA directly uses the pretrained video model's diffusion objective to guide geometry compression, producing geometry latents that are more compatible with video generation. The aligned geometry space is then used for joint video–geometry modeling and further distilled with dual-stream distribution matching for efficient generation. Together with the **GeoVid-80K** dataset, GeoLA achieves faster joint training and improved geometric consistency while maintaining strong visual quality. These results demonstrate that implicit geometry-video joint modeling provides an effective and scalable path toward world-consistent video generation.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *Advances in Neural Information Processing Systems*, 38:57602–57622, 2026.

Aritra Bhowmik, Denis Korzhenkov, Cees GM Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. *arXiv preprint arXiv:2510.19022*, 2025.

Kwon Byung-Ki, Qi Dai, Lee Hyoseok, Chong Luo, and Tae-Hyun Oh. Jointdit: Enhancing rgb-depth joint modeling with diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25261–25271, 2025.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

Yabo Chen, Yuanzhi Liang, Jiepeng Wang, Tingxi Chen, Junfei Cheng, Zixiao Gu, Yuyang Huang, Zicheng Jiang, Wei Li, Tian Li, et al. Teleworld: Towards dynamic multimodal synthesis with a 4d world model. *CoRR*, 2025.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *ICPR*, 2010.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25358–25369, 2026a.

Kaiyi Huang, Yukun Huang, Yu Li, Jianhong Bai, Xintao Wang, Zinan Lin, Xuefei Ning, Jiwen Yu, Pengfei Wan, Yu Wang, et al. Cinescene: Implicit 3d as effective scene representation for cinematic video generation. *arXiv preprint arXiv:2602.06959*, 2026b.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *SIGGRAPH*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *CoRR*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M Rehg. How much 3d do video foundation models encode? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026c.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, 2024.

Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Shubo Lin, Xuanyang Zhang, Wei Cheng, Weiming Hu, Gang Yu, and Jin Gao. Mmphysvideo: Scaling physical plausibility in video generation via joint multimodal modeling. *arXiv preprint arXiv:2604.02817*, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27326–27337, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *CoRR*, 2025.

Boming Tan, Xiangdong Zhang, Ning Liao, Yuqing Zhang, Shaofeng Zhang, Xue Yang, Qi Fan, and Yanyong Zhang. Dreamworld: Unified world modeling in video generation. *arXiv preprint arXiv:2603.00466*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025c.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, et al. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10915–10923, 2026.

Xunzhi Xiang, Yabo Chen, Guiyu Zhang, Zhongyu Wang, Zhe Gao, Quanming Xiang, Gonghu Shang, Junqi Liu, Haibin Huang, Yang Gao, et al. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation. *CoRR*, 2025.

Mingyang Xie, Numair Khan, Tianfu Wang, Naina Dhingra, Seonghyeon Nam, Haitao Yang, Zhuo Hui, Christopher Metzler, Andrea Vedaldi, Hamed Pirsiavash, et al. Lavr: Scene latent conditioned generative video trajectory re-rendering using large 4d reconstruction models. *arXiv preprint arXiv:2601.14674*, 2026.

Haotian Xue, Qi Chen, Zhonghao Wang, Xun Huang, Eli Shechtman, Jinrong Xie, and Yongxin Chen. Mogan: Improving motion quality in video diffusion via few-step motion adversarial post-training. *arXiv preprint arXiv:2511.21592*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

8 APPENDIX

8.1 PRELIMINARY

Diffusion models. Diffusion models (Ho et al., 2020; Song et al., 2020; Peebles & Xie, 2023) generate data by transforming Gaussian noise into samples from the target distribution, and have been widely used in image synthesis (Rombach et al., 2022; Yin et al., 2024), multi-view generation (Mildenhall et al., 2021; He et al., 2024; Kerbl et al., 2023; Bai et al., 2025a), and video generation (Guo et al., 2024). Given clean data $\mathbf{x}$, the forward noising process is

$$\mathbf{x}_t = F(\mathbf{x}, \boldsymbol{\epsilon}, t) = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

where α_t and σ_t control the signal-to-noise ratio. A denoising model can be trained to predict the clean sample by

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{\mathbf{x},t,\boldsymbol{\epsilon}} \left[\|\mathbf{x} - \hat{\mathbf{x}}_\theta(\mathbf{x}_t, t)\|_2^2 \right]. \tag{9}$$

Alternative parameterizations predict the noise $\boldsymbol{\epsilon}$ or velocity $\mathbf{v}$. All parameterizations can be converted into a denoised estimate $\boldsymbol{\mu}(\mathbf{x}_t, t)$, giving the score estimate

$$s(\mathbf{x}_t, t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = -\frac{\mathbf{x}_t - \alpha_t \boldsymbol{\mu}(\mathbf{x}_t, t)}{\sigma_t^2}. \tag{10}$$

Geometry foundation model. Geometry foundation models encode spatial layouts and scene structures from input videos. Given frames $\mathbf{I} \in \mathbb{R}^{B \times T_g \times H_I \times W_I \times 3}$, a geometry encoder $\mathcal{E}_g$ extracts a multi-level token hierarchy $\{\mathbf{G}^\ell\}_{\ell=1}^L = \mathcal{E}_g(\mathbf{I})$. Although these tokens can be decoded into explicit geometry such as dense depth, camera poses, or point clouds, we use them as continuous latent representations of scene structure. This allows geometric knowledge to regularize video generation without exposing the generator to pixel-aligned reconstruction artifacts from decoded geometry. We therefore bypass explicit geometry decoding and directly extract a dense geometry feature by

$$\mathbf{G} = \Gamma\big(\{\mathbf{G}^\ell\}_{\ell=1}^L\big) \in \mathbb{R}^{B \times T_g \times C_G \times H_g \times W_g}. \tag{11}$$

Here, Γ selects the feature from one level of the multi-level geometry representations.

8.2 ARCHITECTURE DETAILS

Backbones. We use Wan2.1-1.3B (Wan et al., 2025) as the backbone of GeoLA throughout all three training stages. For Stage 3 distillation, we additionally train a Wan2.1-14B (Wan et al., 2025) model on Stages 1 and 2 using the same objectives as the 1.3B model. The resulting 14B model then serves as the teacher, while the 1.3B model obtained after Stage 2 is used as the student.

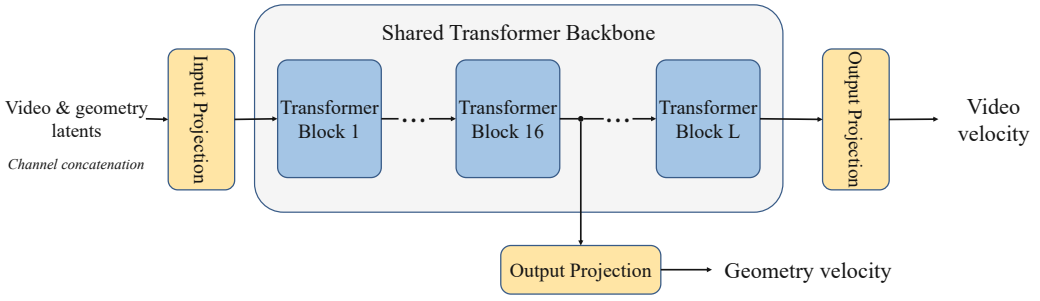


Figure 8: **Architecture of the joint DiT model.** Joint modeling is enabled solely by modifying existing linear projection layers and inserting additional linear projection layers, while keeping the rest of the DiT backbone unchanged.

As illustrated in Fig. 8, We retain the original transformer blocks and modify only the input and output projections to support joint video–geometry modeling. The geometry and video latents are then concatenated along the channel dimension and passed through the input projection. The output projections predict the video and geometry velocities for joint flow matching. These modifications allow the two modalities to share the same transformer backbone without changing its attention or feed-forward layers.

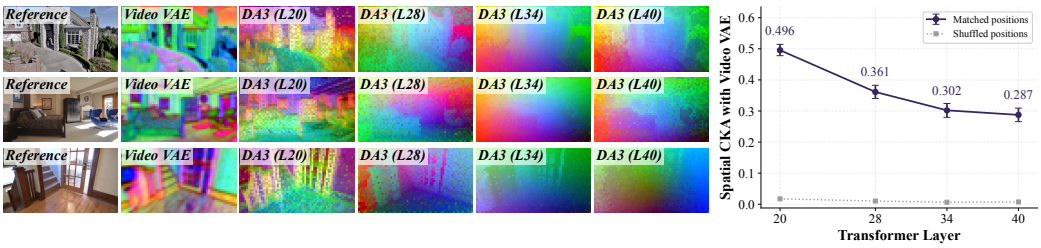


Figure 9: **Selecting geometry features for video–geometry joint modeling.** (a) PCA visualizations of paired video VAE latents and four DA3-Giant feature levels. L20 preserves clearer local scene boundaries, whereas deeper levels exhibit smoother, coarser spatial patterns. (b) Spatial CKA with video VAE latents on paired videos at a matched spatial resolution. L20 achieves the strongest spatial correspondence. Error bars denote 95% bootstrap confidence intervals over videos; the dotted line shows a spatially shuffled control.

Geometry feature extraction. DepthAnything3-Giant (Lin et al., 2025) provides intermediate features at four transformer layers: 20, 28, 34, and 40. We use the features from layer 20, i.e., the first level, as input to the geometry adapter. As shown in Fig. 9(a), the first-level features retain sharper local scene boundaries, while deeper features become progressively smoother and spatially coarser. This observation is supported by Centered Kernel Alignment (CKA), which measures the similarity between feature representations. At matched spatial resolutions, the first level shows the highest spatial CKA with video VAE latents (0.496), compared with 0.361, 0.302, and 0.287 for the three deeper levels (Fig. 9(b)). Edge-map analysis further confirms this trend: the first-level features correlate more strongly with RGB boundaries than the deepest features (0.339 vs. 0.064). Together, these results motivate our use of the first-level features for geometry–video alignment.

8.3 Training Details

Training schedule. We train for 4,000 iterations in Stage 1, 20,000 iterations in Stage 2, and 2,000 iterations in Stage 3. We use a global batch size of 48 in all three stages. Training clips contain 81 frames at a resolution of 480×832, corresponding to video latents with 21 temporal positions, 16 channels, and a spatial resolution of 60×104. Table 7 summarizes the optimization settings.

Ablation settings. To ensure a controlled comparison, we adapt both Gen3R (Huang et al., 2026a) and DreamWorld (Tan et al., 2026) based on their respective papers and released implementations. For both methods, we replace the VGGT backbone in the LRM with DepthAnything3-Giant and use geometry features from the same feature level as our method for joint modeling. We also replace sequence-wise feature concatenation with channel-wise concatenation to maintain a comparable computational cost. For Gen3R, we additionally train a geometry reconstruction VAE for 20,000 iterations to compress the geometry features, whereas for DreamWorld, we directly apply PCA for feature compression. For both baselines, we follow the same training schedule as our method, with 4,000 iterations of DiT warm-up followed by 20,000 iterations of joint training. Therefore, the 4,000-step phase is not an additional DiT training budget for GeoLA. It is the pre-joint phase used by all methods. In GeoLA, this phase updates the DiT together with the geometry adapter for a more align latent space. For Gen3R and DreamWorld, the corresponding phase follows their respective warm-up procedures using their compressed geometry representations. All methods then undergo the same 20,000-step joint video–geometry training.

Geometry representation and prediction head. We use 1,536-dimensional features from layer 20 of the frozen DA3-Giant encoder. The cached features contain 21 temporal positions on a 36×64 spatial grid. For each temporal position, the geometry adapter first bilinearly resizes the features to 60×104 and then applies a token-wise lightweight MLP and a GELU activation between the linear layers. This produces four geometry channels on the video-latent grid without additional temporal resampling inside the adapter. Inspired by recent works (Dai et al., 2025; Huang et al., 2026c), we choose an intermediate layer of the DiT, specifically the 16th layer, to predict the

Table 7: Training hyperparameters for the three stages of GeoLA.

Hyperparameter	Stage 1 Geometry adaptation	Stage 2 Joint modeling	Stage 3 Joint distillation
Training iterations	4,000	20,000	2,000
Global batch size	48	48	48
Optimizer	AdamW	AdamW	AdamW
AdamW (β_1, β_2)	$(0, 0.999)$	$(0, 0.999)$	$(0, 0.999)$
AdamW ϵ	10^{-8}	10^{-8}	10^{-8}
Weight decay	0	0	0.01
Learning rate (1.3B / student)	10^{-5}	10^{-5}	10^{-5}
Learning rate (14B / teacher)	10^{-6}	10^{-6}	Frozen
Learning-rate schedule	Constant	Constant	Constant
Fake-score learning rate	–	–	10^{-6}
Fake-score / generator updates	–	–	$5:1$
Generator EMA decay	–	–	0.99
EMA initialization iteration	–	–	200
Video resolution $(H \times W)$	480×832	480×832	480×832
Video / latent temporal length	$81/21$	$81/21$	$81/21$
Mixed precision	bfloat16	bfloat16	bfloat16
Timestep grid size	1,000	1,000	1,000
Timestep shift	8	8	5
Score-evaluation timestep range	–	–	$[20, 980]$
Geometry adapter	Trainable	Frozen	Frozen
DMD loss weights (video, geometry)	–	–	$(1, 0.1)$
Generation steps	50	50	5

geometry flow velocity, while the final output head is used to predict the video flow velocity. The geometry head's output projection is initialized to zero.

Optimization across stages. In Stage 1, we jointly optimize the video DiT and the geometry adapter using the conditional video flow-matching loss. The adapter compresses the geometry features into latents that condition video denoising. Gradients from the video denoising loss update both the adapter and the DiT. Keeping the DiT trainable allows it to adapt to the geometry conditions while the adapter learns to retain information useful for video denoising.

In Stage 2, we freeze the geometry adapter and train the DiT to jointly predict the video and geometry velocities. The DiT is initialized from Stage 1, retaining its adaptation to the geometry inputs. Freezing the adapter keeps the geometry latent space fixed during joint training.

In Stage 3, we use the 14B joint model as teacher model and distill it into the 1.3B student using joint distribution matching. The teacher provides supervision for both video and geometry generation.

8.4 More Results

Abalation on geometry feature dimension We vary the channel dimension Cg of the geometry latent after the adapter projection on image-to-video settings. As shown in Table 8, a small geometry feature dimension is already sufficient for effective joint modeling. This suggests that additional geometry channels can strengthen sparse correspondence-level consistency, but also increase the burden of jointly modeling geometry and video appearance under

Table 8: **Ablation on Geometry Feature Dimension.** We vary the channel dimension C_g of the adapted geometry latent while keeping the remaining pipeline unchanged.

Geometry Dimension	3D Reconstruction Error			3D Consistency Metrics	
	PSNR ↑	SSIM ↑	LPIPS ↓	Epipolar Error ↓	Inlier Rate ↑
$C_g = 4$	**21.18**	**0.728**	**0.311**	5.35	0.778
$C_g = 8$	20.41	0.710	0.323	6.15	0.755
$C_g = 16$	20.19	0.703	0.350	**4.37**	**0.815**



Figure 10: **Point Cloud Reconstruction from GeoLA Videos.** We reconstruct 3D point clouds from videos generated by GeoLA. The results show coherent scene geometry, complete object structures, and consistent cross-view alignment under camera motion.



Figure 11: **3DGS Reconstruction from GeoLA Videos.** We reconstruct videos generated by GeoLA using 3D Gaussian Splatting and visualize the resulting 3D scenes. The reconstructions preserve stable scene layouts and consistent geometry across viewpoints.

limited model capacity. Therefore, we adopt a small geometry feature dimension, which provides the best overall balance between reconstruction fidelity and geometric consistency.

3D reconstruction results. To further evaluate the geometric consistency of the generated videos, we additionally perform 3D reconstruction using point clouds and 3D Gaussian Splatting (3DGS), as shown in Fig. 10 and Fig. 11. The reconstructed point clouds recover the overall scene structure with reasonably coherent geometry, while the 3DGS reconstructions preserve the main scene layout and visual appearance with limited artifacts. As these reconstruction methods rely on cross-frame correspondences and multi-view consistency, the resulting 3D representations suggest that our generated videos maintain a relatively stable underlying geometry across frames.

8.5 Problem Analysis Details

We compare two compression architectures, each trained with or without video-latent alignment supervision. All variants use the same input geometry features and produce four-channel latents on an identical spatiotemporal grid. Let X and $\hat{X}$ denote the input and reconstructed feature tensors. The reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{P}|C} \sum_{p \in \mathcal{P}} \|\hat{x}_p - x_p\|_2^2, \tag{12}$$

where C is the input feature dimensionality and $\mathcal{P}$ indexes spatiotemporal positions across the batch.

PCA. We use a linear autoencoder initialized by global PCA estimated from a shared training subset. Encoding and reconstruction share an orthonormal basis $U \in \mathbb{R}^{C \times 4}$:

$$z_p = U^\top (x_p - \mu), \qquad \hat{x}_p = U z_p + \mu, \qquad U^\top U = I_4. \tag{13}$$

The feature mean μ remains fixed. We optimize U with $\mathcal{L}_{\text{rec}}$ while maintaining its orthonormality through QR parameterization. The projection is shared across samples, and the decoder introduces no independent trainable parameters.

PCA + KL. We retain the same linear architecture and PCA initialization and add supervision from paired video VAE latents. A trainable pointwise affine projection $P : \mathbb{R}^4 \to \mathbb{R}^{16}$ matches the geometry and video latent channel dimensions. The alignment objective is

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_p[D_{\text{KL}}(\sigma_c(v_p) \,\|\, \sigma_c(P(z_p)))] \,, \tag{14}$$

where v_p is the fixed video VAE latent and σ_c denotes softmax across channels. Training minimizes $\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$. The auxiliary projection is discarded after pretraining.

Rec. We replace the linear compression model with a deterministic convolutional autoencoder based on the Wan VAE architecture:

$$Z = E_\theta(X), \qquad \hat{X} = D_\phi(Z), \qquad \min_{\theta,\phi} \mathcal{L}_{\text{rec}}. \tag{15}$$

The encoder and decoder are jointly trained to reconstruct the input geometry features through the four-channel bottleneck.

Rec. + KL. We retain the same convolutional encoder and decoder and introduce the same auxiliary projection and KL objective defined above. Training minimizes $\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}$, enabling comparison with **Rec.** under a fixed compression architecture.

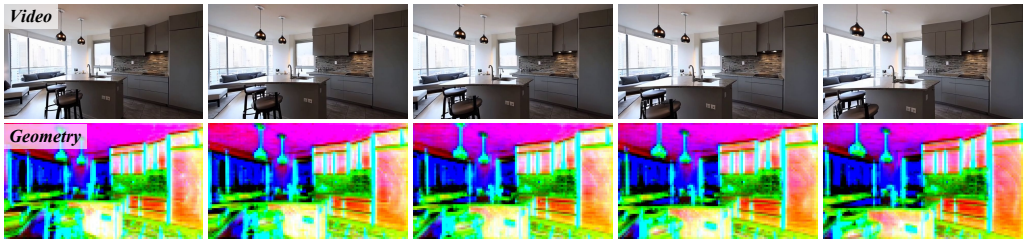
After pretraining, each compression encoder is frozen. Its latents are standardized using its own fixed per-channel mean and standard deviation, estimated on the same training subset across variants. All variants then follow the same conditional and joint video–geometry training protocol.

8.6 LLM Usage

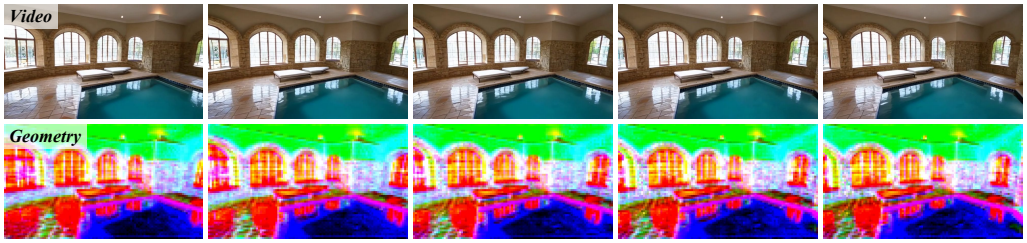
Large language models (LLMs) were used to assist with the writing and polishing of the manuscript, primarily to improve readability, grammatical correctness, and clarity of expression. Their use was limited to language refinement and did not contribute to the development of the research methodology, experimental design, data analysis, or conclusions. All technical content and research findings were produced and verified by the authors, who take full responsibility for the final manuscript.



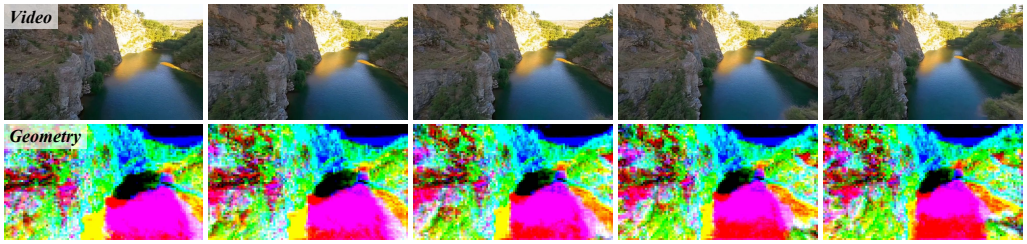
Text: Strong camera movement, dynamic shot. A pastel nursery with a white crib, animal murals, soft rugs, a rocking chair, stuffed toys, and hanging cloud decorations; the room remains still while the camera floats in from the doorway, circles the crib, rises toward the ceiling decorations, tilts down over the rug, and glides toward the window.



Text: Strong camera movement, dynamic shot. A compact modern apartment kitchen connected to a living room, with matte black cabinets, stone backsplash, bar stools, pendant lights, and city windows; the interior remains static while the camera tracks from the sink across the island, curves into the living area, rises over the sofa, pans to the skyline, and pulls back.

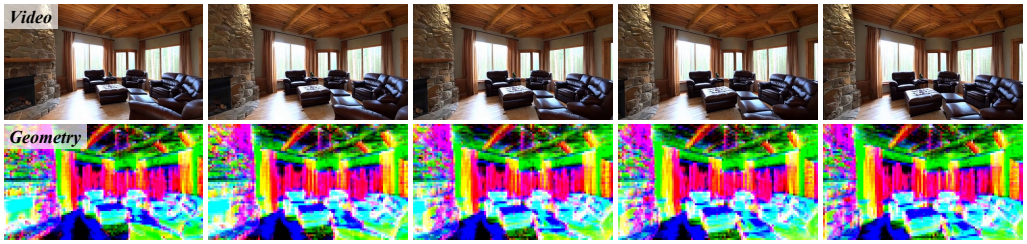


Text: Strong camera movement, dynamic shot. A spacious indoor pool room with arched windows, stone columns, tiled floors, lounge chairs, and blue water reflections; the room remains still while the camera cranes down from a high corner, flies along the pool surface, sweeps around a column, rises toward the windows, and pulls back diagonally.

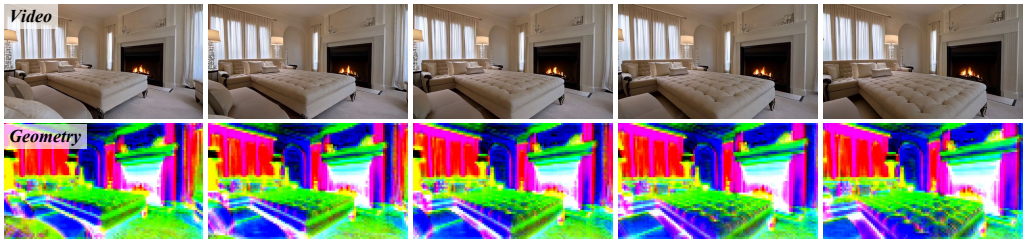


Text: Strong camera movement, dynamic shot. A rocky river gorge with clear water between cliffs, pine trees along the edges, and sunlight striking the stone; the landscape remains static while the camera flies low above the river, follows its bend, rises along the cliff wall, arcs over the gorge, and pulls into a high view.

Figure 12: **Additional qualitative results of paired video and geometry generation.** Each example shows the jointly generated video and its corresponding geometry output.



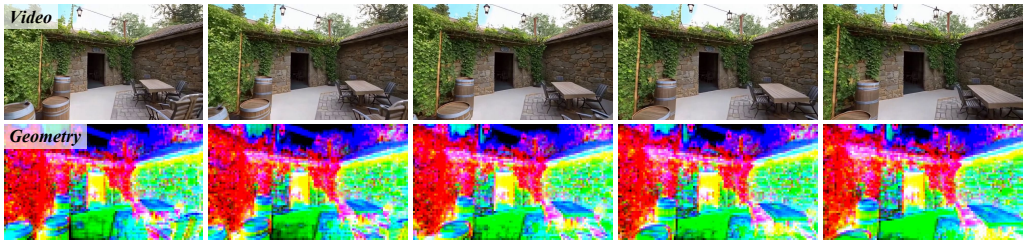
Text: Strong camera movement, dynamic shot. A cozy rustic living room with a towering stone fireplace, dark leather recliners, a matching ottoman, wooden ceiling beams, brown curtains, and tall windows facing a pine forest; the room remains still while the camera starts outside the window, pushes inward, sweeps in a wide arc around the seating area, cranes upward to the beams, and pulls back toward the fireplace.



Text: Strong camera movement, dynamic shot. A luxury master suite with a fireplace, cream bedding, upholstered bench, crystal lamps, and balcony doors with sheer curtains; the room remains still while the camera pushes from the sitting area toward the bed, orbits the bench, rises above the pillows, pans toward the balcony, and pulls back.



Text: Strong camera movement, dynamic shot. A desert villa with white walls, cactus plants, terracotta paving, a blue reflecting pool, and distant mountains under bright sunlight; the exterior remains static while the camera flies over the courtyard wall, circles the pool, glides through the open terrace, rises above the building, and tilts down over the full layout.



Text: Strong camera movement, dynamic shot. A vineyard courtyard with stone walls, wooden barrels, climbing vines, outdoor tables, and warm lanterns; the courtyard remains still while the camera moves through the entrance, circles a barrel table, rises over the vines, pans toward the vineyard rows, and pulls backward through the archway.

Figure 13: **Additional qualitative results of paired video and geometry generation.** Each example shows the jointly generated video and its corresponding geometry output.